WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Natural Language Processing (NLP) from Image to Text

[1]Isha Sudhir Sawant

[1]Student, Masters of Computer Applications, Jain (Deemed-To-Be-University), Bangalore, India,
[1]ishusawant752@gmail.com
DOI: https://doi.org/10.55248/gengpi.5.0524.1463

A B S T R A C T

The rapid advancements in Natural Language Processing (NLP) have catalyzed significant breakthroughs in field of image-to-text conversion. Enabling machines to understand and describe visual content with increasing accuracy. This paper explores the intersection of computer vision and NLP. It focuses on methodologies and applications of translating images into coherent textual descriptions. We delve into various approaches employed in this domain. Including convolutional neural networks (CNNs) for image feature extraction. Recurrent neural networks (RNNs) for sequence generation. And transformer models that integrate both functionalities for enhanced performance.

The paper also examines challenges inherent in this task. Such as handling diverse and complex visual scenes. Generating contextually relevant and grammatically correct descriptions ensuring cultural and contextual sensitivity. Additionally, we discuss practical applications of image-to-text technology. In areas such as assistive technology for visually impaired automated content creation and enhanced human-computer interaction.

By reviewing state-of-the-art models and their performance on benchmark datasets. This paper aims to provide comprehensive overview of current landscape and future directions. In the field of NLP from image to text.

Keywords: Natural Language Processing (NLP), Image to Text, Optical Character Recognition (OCR), Image Captioning, Computer Vision, Deep Learning, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Transformers.

1. INTRODUCTION

Natural Language Processing (NLP) is branch of artificial intelligence. It focuses on enabling computers to interact with humans using natural language. It encompasses development of algorithms. Models help machines understand interpret and generate human language. Integrating NLP with computer vision, particularly in transforming images into text marks a significant advancement. It facilitates applications such as automatic image captioning. Scene understanding and visual question answering are also enhanced. This research paper explores detailed methodologies. Applications and challenges of NLP from image to text will be examined.

Topics for the Research Paper

Overview of Natural Language Processing (NLP)

Definition and Scope: Understanding what NLP is and its boundaries within AI.

Historical Background: Tracing evolution of NLP from inception to modern-day advancements

Core Components: Key techniques such as syntax (structure). Semantics (meaning). Pragmatics (context).

Introduction to Computer Vision

Definition and Scope: Defining computer vision and its objectives

Historical Context: Milestones and significant breakthroughs in computer vision

Fundamental Techniques: Basics of image processing. Object detection and image segmentation

Integration of NLP and Computer Vision

Multi-modal AI: Combining NLP and computer vision for enhanced capabilities

Importance: necessity and benefits of integrating these technologies. Overview of Image-to-Text Basic concepts. Processes involved in generating text from images

Image Captioning

Definition and Significance: Image captioning critical task in AI, involves automatically generating textual descriptions of images. Its importance lies in facilitating accessibility for visually impaired individuals. Also enhancing content-based image retrieval systems.

Key Methodologies: Techniques. Encoder-decoder frameworks. CNNs. RNNs. Each has unique benefits in handling complex visual and linguistic data. Encoder-decoder frameworks, particularly transformative in mapping visual inputs to textual outputs. CNNs specialize in extracting spatial features from images. Whereas RNNs excel in sequence prediction.

State-of-the-Art Models: Representing pinnacle of research, examples include Show and Tell. Also Neural Image Caption Generator. These models integrate various sophisticated techniques. They demonstrate remarkable efficacy in generating coherent, contextually relevant descriptions.

Scene Text Recognition

Definition and Applications: Identifying and using text within images. The extraction of textual content from images known as Optical Character Recognition (OCR), facilitates numerous applications. These include document digitization license plate recognition and automated data entry systems. The widespread usage spans industries from healthcare to finance. Enabling enhanced data accessibility and operational efficiency.

Techniques: Methods for detecting and recognizing text in different conditions are critical for accuracy. Machine learning algorithms have significantly advanced text recognition capabilities. Traditional techniques relied on template matching and edge detection. Modern approaches include convolutional neural networks (CNNs), which excel at processing complex visual data. Pre-processing steps like binarization and noise reduction improve quality. Text detection is usually followed by text recognition which transforms identified regions into machine-readable characters.

Challenges: Issues inherent to text recognition systems include font variability and different orientations. Various fonts and styles can confuse OCR systems. Additionally text alignment and lighting effects pose significant problems. Skewed or rotated text often requires corrective algorithms. Similarly, inconsistent lighting can obscure characters. Overcoming these challenges necessitates improved models and robust preprocessing techniques. Leveraging extensive datasets for training can mitigate some of these issues though variability remains a persistent obstacle.

Visual Question Answering (VQA)

Overview: Understanding VQA and its practical uses. Visual Question Answering (VQA) stands at intersection of computer vision and natural language processing. It aims to create systems capable of answering questions about images. VQA's practical uses are far-reaching. They find applications in fields like autonomous driving medical imaging and interactive systems.

Architectures: Common system designs used in VQA. Several architectures have been proposed for VQA systems. Two primary components underpin them. A vision model processes image input. A language model interprets the question. These components fuse information, arriving at an answer. The architecture selection significantly influences system's performance and effectiveness. Modern advancements such as attention mechanisms, have further enhanced these models.

Datasets and Evaluation: Resources and methods for assessing VQA systems. Datasets are vital for training validating and testing these systems. Popular datasets include VQA, COCO-QA and VizWiz. Each catering to different aspects of VQA tasks. Evaluation metrics include accuracy, mean reciprocal rank and specific to VQA language-based metrics assessing clarity and relevance. However, challenges persist. Dataset biases and varying question complexities affect the generalizability of these models.

Challenges in NLP from Image to Text

Technical Challenges: Addressing ambiguities. Contextual complexities.

Dataset Limitations: Issues related to data quality. Bias remains concern.

Scalability: Managing computational demands. Efficiency is also crucial.

Applications and Use Cases

Assistive Technologies: Tools aiding visually impaired.

Content Creation: Automating generation and management of content.

Autonomous Systems: Enhancing robotics. Self-driving technologies.

Educational Tools: Interactive adaptive learning systems.

Future Directions. Research Opportunities

Emerging Trends: New developments combining NLP and computer vision.

Potential Advancements: Improvements algorithms model performance.

Interdisciplinary Research: Exploring areas for cross-disciplinary innovation.

2.LITERATURE REVIEW

The integration of Natural Language Processing (NLP) with computer vision has garnered significant attention in recent years due to its potential. It bridges gap between visual and textual data. This interdisciplinary approach has led to advancements. Notably in applications such as image captioning. Scene text recognition and visual question answering (VQA). The surge in deep learning techniques has further propelled progress in this domain. Enabling more accurate and sophisticated models.

Recent studies have explored various methodologies to enhance performance and robustness of models that convert images to text. These methodologies include advanced neural network architectures. And attention mechanisms. As well as multi-modal learning frameworks. The focus has shifted Developing models that not only generate coherent and contextually relevant captions. Also understand complex scenes

This literature review examines recent contributions from past five years (2019-2023). It provides comprehensive overview of state-of-the-art techniques in NLP from image to text. Selected papers from IEEE journals highlight significant advancements challenges. Future directions in this field are also discussed.

The following table summarizes key papers. It presents titles. Authors' publication years and journal names and brief summary of contribution ***Tables***

All tables should be numbered with Arabic numerals. Every table should have a caption. Headings should be placed above tables, left justified. Only horizontal lines should be used within a table, to distinguish the column headings from the body of the table, and immediately above and below the table. Tables must be embedded into the text and not supplied separately. Below is an example which the authors may find useful.

Literature Review Table

Title	Authors	Year	Journal	Summary
Deep Learning-Based Image Captioning with Improved GRU Models	Saeed Anwar, Salman Khan, Nick Barnes	2019	IEEE Access	Improved image captioning results by integrating advanced GRU models with visual attributes.
Hybrid Attention Networks for Visual Question Answering	Gao, Peng, et al.	2019	IEEE Transactions on Image Processing	Utilized hybrid attention networks to improve VQA through focusing on crucial image regions and questions.
Unified Vision-Language Pre-Training for Image Captioning and VQA	Zhe Gan, Linjie Li, Chunyuan Li, et al.	2020	IEEE Transactions on Pattern Analysis and Machine Intelligence	Suggested a single vision-language pre-training model that improves performance on both image captioning and VQA assignments.
Scene Text Recognition with Attention Mechanisms	Baoguang Shi, Xiang Bai, Cong Yao	2020	IEEE Transactions on Pattern Analysis and Machine Intelligence	Created a model that focuses on attention to identify text in real-world images, enhancing accuracy and resilience.
End-to-End Dense Video Captioning with Masked Transformer	Luowei Zhou, Junjie Xu, Licheng Yu, et al.	2020	IEEE Transactions on Pattern Analysis and Machine Intelligence	Developed a masked transformer model to produce detailed and contextually relevant captions for videos.
Multi-Modal Transformer for Image Captioning	Jaemin Cho, Kyoung-Woon On, Jooyoung Kim, et al.	2021	IEEE Transactions on Image Processing	Suggested a transformer model that utilizes both visual and textual information to improve image captioning.
Visual Question Answering with Dynamic Attention and External Knowledge	Xinxin Zhu, Jingjing Liu, Hanwang Zhang	2021	IEEE Transactions on Neural Networks and Learning Systems	Studied how utilizing dynamic attention and external knowledge can enhance performance in VQA.
Learning to Caption Images with Object Detection and Fine-Grained Attributes	Peng Wang, Qi Wu, Chunhua Shen, et al.	2021	IEEE Transactions on Pattern Analysis and Machine Intelligence	Created a framework that combines object detection with detailed attribute data to improve image captioning.
Robust Scene Text Recognition with Semantic Segmentation	Zhanzhan Cheng, Fan Bai, Yunlu Xu, et al.	2022	IEEE Transactions on Image Processing	Improved recognition of text in scenes by integrating semantic segmentation

				methods to address intricate backgrounds.
Dense Captioning Transformers for Visual Question Answering	Chen, Long, et al.	2022	IEEE Transactions on Neural Networks and Learning Systems	Introduced dense captioning transformers that enhance VQA through the creation of in-depth image descriptions.
Image Captioning with Visual-Semantic LSTM and External Memory	Wang, Yufei, et al.	2022	IEEE Transactions on Multimedia	Presented a visual-semantic LSTM model incorporating external memory to enhance context comprehension in image captioning.
Vision-Language Navigation with Pre-Trained Multimodal Representations	Hao Tan, Mohit Bansal	2023	IEEE Transactions on Pattern Analysis and Machine Intelligence	Suggested a vision-language navigation model that utilizes pre-trained multimodal representations to enhance performance.
Hierarchical Attention Networks for Visual Question Answering	Kang, Donghyun, et al.	2023	IEEE Transactions on Image Processing	Created hierarchical attention networks to improve VQA by capturing contextual information at multiple levels.
Adaptive Attention Networks for Image Captioning and VQA	Lu, Jiasen, et al.	2023	IEEE Transactions on Pattern Analysis and Machine Intelligence	Suggested adaptive attention networks that modify focus in real-time to improve results in image captioning and VQA tasks.
Self-Supervised Learning for Scene Text Recognition	Tian, Zhi, et al.	2023	IEEE Transactions on Pattern Analysis and Machine Intelligence	Investigated self-supervised learning methods in order to enhance the precision and robustness of scene text recognition models.

3.METHODS OR TOOLS OR ALGORITHMS

***Methods*:**

Optical Character Recognition (OCR) :- Traditional OCR: Uses pattern recognition and matrix matching to convert printed text in images to machine-encoded text. Modern OCR: Utilizes deep learning techniques. Such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for more accurate text extraction.

Scene Text Recognition (STR):- Text Proposal Networks: Methods like connectionist text proposal networks (CTPN) identify text regions in natural scene images. End-to-End Models: Combine text detection. Recognition in single model using deep learning frameworks

Handwritten Text Recognition (HTR):- Recurrent Neural Networks (RNNs): Used for sequence prediction. Often combined with LSTM (long short-term memory) networks To handle varying writing styles. Transformers: Recent advancements in using transformer architectures. Improve recognition accuracy and handle long-range dependencies in handwritten text.

Tools:

Tesseract OCR :- Description Tesseract is open-source OCR engine. It was developed by Google. It is highly popular for its accuracy. It supports multiple languages. Features: Supports more than 100 languages. Can be trained. For custom fonts and languages. Provides output in plain text. hOCR (HTML for OCR) Searchable PDF formats. Usage. Suitable for extracting text. From printed documents. Books. And articles.

Google Cloud Vision API :- Description: Cloud-based OCR tool leverages Google's machine learning technology. It extracts text from images

Advantages: Cloud-based OCR systems offer multiple advantages Firstly they can handle large volumes. Secondly, they provide accurate and fast text extraction. Thirdly they integrate seamlessly with other applications. Disadvantages: There are some disadvantages. Data privacy remains a significant concern. Additionally, reliance on internet connectivity can hinder performance. Furthermore these solutions might incur ongoing costs.

Amazon Textract:- Description: Service by Amazon Web Services (AWS) that automatically extracts text. Data from scanned documents.

Functionality Overview: Amazon Textract is innovative solution for extracting text and data from scanned documents. This service is designed to recognize various text and data formats. It accurately processes printed text. Handwriting and checkboxes. Furthermore. It supports diverse document types such as forms tables and contracts. This becomes invaluable in reducing the time. And labor previously required for manual data entry.

Capabilities and Features: Textract employs machine learning models trained continuously to improve extraction accuracy. These models enable Textract to handle complex documents. Automatically identifying key-value pairs and nested data. For example it can distinguish between invoice numbers payment terms, total amounts and line item details on single document. Additionally it uses APIs that allow integration. Various software systems integration enabling seamless workflow automation.

PyTorch and TensorFlow:- TensorFlow and PyTorch Comprehensive Comparison. TensorFlow and PyTorch which are two dominant deep learning frameworks have become indispensable tools for researchers and engineers. Both originated from tech giants. TensorFlow was developed by Google Brain. PyTorch was developed by Facebook's AI Research lab.

Performance and Speed: Optimizing computational efficiency and speed constitutes crucial aspects of deep learning frameworks. TensorFlow often excels in performance due to advanced optimizations. And XLA (Accelerated Linear Algebra) In contrast PyTorch focuses on providing intuitive dynamic computational graph. Thereby facilitating a more flexible approach. This results in slightly slower performance in some scenarios.

Usability: The variability in syntax and usability between TensorFlow and PyTorch raises preferences among users. TensorFlow possesses higher complexity. Engaging users in learning curve. Alternatively PyTorch offers more straightforward syntax. Appealing to both beginners and experienced developers many advocates prefer PyTorch. They argue it aligns closely with Python’s innate programming paradigms. Thereby streamlining workflow.

TensorFlow: Developed by Google Brain TensorFlow evolved from inception in 2015. It is designed to facilitate implementation. And deployment of machine learning models. TensorFlow is renowned for scalability and flexibility. Such attributes make it suitable for production-grade applications. Its ecosystem is diverse. Encompassing TensorFlow Extended (TFX) for deploying and managing models at scale. TensorFlow Lite for mobile. And embedded device computations

PyTorch: PyTorch offering from Facebook's AI Research lab (FAIR) emphasizes simplicity and dynamic computational graphing. Introduced 2016. PyTorch enables rapid prototyping and facilitates experimentation. Its dynamic nature allows computational graphs to be altered during runtime. This feature significantly simplifies debugging. Popular among researchers. PyTorch's integration with native Python proves advantageous. Support for custom extensions due to its CUDA integration

OpenCV:Description: Open-source computer vision library provides tools for image. Also video processing. Key Features: Offers extensive collection of algorithms. Supports multi-platform usage. Includes pre-trained neural networks for deep learning. Integrates seamlessly with other libraries. It facilitates real-time. Processing. Documentation: Comprehensive. Clearly outlines functionality installation process and code examples.

Keras-OCR:- Description: High-level OCR library built on Keras and TensorFlow. Simplifies development of OCR models within a comprehensive framework. Library provides pre-trained models which can be easily fine-tuned for specific tasks. This flexibility makes it particularly suited for research and industrial applications. Compatibility is key strength. This library seamlessly integrates with other popular machine learning tools such as scikit-learn. Thus enabling interoperability. Easing model deployment.

LaTeX:- Description: Document preparation system used for high-quality typesetting. Often used research papers. System Overview: LaTeX is widely adopted sophisticated typesetting system. Provides powerful tools. Its ability to handle large documents makes it invaluable. For academic, mathematical and scientific writing. Employs markup language. Commands are embedded directly within text. Rather than relying on GUI. This allows precise control over formatting.

Algorithms:

Tesseract OCR:- Overview: Tesseract is open-source OCR engine developed by Google. It can recognize more than 100 languages. It is widely used due to its flexibility. It supports various scripts.

Preprocessing: Involves converting images to binary. Detecting edges and removing noise. Segmentation: Divides text into lines. Words. Recognition: Uses LSTM networks to recognize text characters. Postprocessing: Includes dictionary look-up to improve accuracy. Correct errors. Applications: Document digitization automated data entry, text extraction from scanned documents.

Convolutional Recurrent Neural Network (CRNN):- Overview: CRNNs combine convolutional neural networks (CNNs) for feature extraction. Recurrent neural networks (RNNs) for sequence prediction making them ideal for text recognition tasks. Architecture: CNN Layer: Extracts features from input image capturing spatial hierarchies. RNN Layer. Processes the extracted features as sequences. Handles varying text lengths and dependencies. CTC (Connectionist Temporal Classification) Loss. Used to train network for variable-length sequences without requiring pre-segmented data. Applications: Scene text recognition. License plate recognition. Handwriting recognition.

Connectionist Text Proposal Network (CTPN):- Overview: CTPN is used for detecting text lines in natural images. It combines text detection. It also combines sequence prediction in single model.

Proposal Generation: Generates text proposals using small sliding window. RNN Integration: Incorporates RNNs Predict text line positions. And connections. Bounding Box Regression: Refines position. And size of detected text lines. Applications: Detecting and recognizing text in natural scenes street signs, billboards.

Attention Mechanisms:- Overview: Attention mechanisms improve performance of OCR and STR models by focusing on relevant parts of image during the recognition process. Attention mechanisms integral to modern OCR (Optical Character Recognition) and STR (Scene Text Recognition) frameworks, revolutionize text recognition. They enable models to selectively concentrate on pertinent segments of an input image. This mimic human reading patterns. Traditional models often struggle with variable text orientations within images. Consequently recognition accuracy suffers.

Mechanism: The core idea of attention involves assigning different weights to parts of input data. The model's architecture incorporates an attention layer. It evaluates relevance of each segment. This is achieved through a learned scoring function. Over iterations, the model gets trained to prioritize impactful information. This leads to enhanced performance. Transformers:-Overview: Transformers originally designed for NLP tasks have been adapted for image-to-text tasks. Their ability to handle long-range dependencies and parallelize computation.

Feature Pyramid Network (FPN) extracts features at different scales. Non-Max Suppression (NMS) removes redundant text proposals. It refines detection results. Quadrilateral Detection detects text regions as quadrilaterals. This improves flexibility in handling various text orientations. Applications include real-time text detection in videos and augmented reality applications. It also includes mobile applications.

4.CONCLUSION & FUTURE

Conclusion:

Optical Character Recognition (OCR): Traditional OCR methods such as pattern recognition matrix matching have largely been supplanted. Modern deep learning techniques have taken precedence. These methods include convolutional neural networks (CNNs). Recurrent neural networks (RNNs) are included as well Tesseract OCR exemplifies such advancements. They have significantly improved text recognition accuracy. This especially true for printed text.

Scene Text Recognition (STR): Algorithms like Connectionist Text Proposal Network (CTPN) and CRNNs effectively detect text. This occurs in natural scenes They excel at recognizing text. Useful for applications requiring real-time processing. Complex backgrounds handled efficiently.

Handwritten Text Recognition (HTR): Using RNNs LSTMs transformers. Modern HTR systems handle variability. They manage complexity This refers to handwritten text. This progress crucial for digitizing handwritten documents. It also enhances accessibility.

Advanced Techniques: Attention mechanisms and transformers further refined text recognition capabilities. This allows models to focus on relevant parts of images. Models handle long-range dependencies more efficiently. These techniques shown significant improvements in both OCR. And STR tasks. Applications and Future Directions Discussed algorithms have broad applications including document digitization. Assistive technology automated data entry. Real-time text recognition in various environments. Future research may focus on overcoming current limitations. Handling low-quality images. Multilingual text recognition. Integrating models with other AI technologies for more robust solutions

Final Thoughts Advancements in NLP from image to text have made significant strides in accuracy. Efficiency and applicability. By continuing to refine these methods field can address remaining challenges. Researchers can explore new technologies. This will expand impact across various domains All authors are required to complete the Procedia exclusive license transfer agreement before the article can be published, which they can do online. This transfer agreement enables Elsevier to protect the copyrighted material for the authors, but does not relinquish the authors' proprietary rights. The copyright transfer covers the exclusive rights to reproduce and distribute the article, including reprints, photographic reproductions, microfilm or any other reproductions of similar nature and translations. Authors are responsible for obtaining from the copyright holder, the permission to reproduce any figures for which copyright exists.

Future:

Multi-Modal Integration: Explore techniques to integrate information from multiple modalities. Text image audio. Integrate for robust and accurate text recognition. Investigate incorporating contextual information from diverse sources. This can improve understanding. And interpretation of the text.

Self-Supervised Learning: Explore self-supervised learning techniques to leverage large amounts of unlabeled data for pretraining text recognition models. Investigate methods like contrastive learning. Pretext tasks to learn representations that capture meaningful semantic information from images and text.

Adversarial Robustness: Develop models that are robust to adversarial attacks. Ensuring reliable performance in real-world scenarios with varying image qualities and potential adversarial manipulations Investigate techniques such as adversarial training. And robust optimization to enhance model resilience against perturbations

Continual Learning: Investigate continual learning strategies. Enable models to adapt to new domains and languages over time without catastrophic forgetting. Explore techniques such as lifelong learning. Meta-learning to efficiently leverage past experiences. Adapt to evolving data distributions.

Ethical Considerations: Address ethical considerations surrounding privacy bias and fairness in image-to-text systems. The development and deployment of image-to-text systems demand rigorous attention to ethical considerations. These considerations pertain notably to the triad of privacy, bias and fairness They are paramount in ensuring these systems' responsible usage. The vast amounts of visual data processed by such systems necessitate stringent privacy protocols. This is to protect individuals' confidential information adequately. Bias in training data constitutes a significant challenge. It perpetuates inequities across various demographic groups. Techniques to identify and rectify these biases are essential. This allows for an inclusive and fair technological environment. Researchers must continually refine methods to minimize these biases. They should employ diverse datasets that better represent varied user populations. Ensuring fairness in the outcomes of image-to-text systems is an ongoing imperative This requires interdisciplinary collaboration. Legal social and technical expertise must converge to create robust frameworks. These frameworks must guided by ethical principles that promote justice and equality.

Develop methods to mitigate biases in training data and ensure equitable outcomes for diverse user populations. Mitigating biases in training data demands comprehensive strategies. These should span from data collection to algorithmic adjustments. One effective approach involves curating diverse datasets This ensures the representation of different genders, races and socio-economic backgrounds. Post-processing techniques also play a critical role. These techniques adjust algorithmic outputs to rectify identified biases. Continuous monitoring and assessment of system performance are crucial. This is to identify and address biases that may emerge during deployment. Engaging with diverse user populations during the development phase is essential. This inclusive approach provides invaluable feedback. It improves system design to equitably serve all users. Establishing transparent practices in algorithm development and deployment further enhances trust and accountability.

References

1. Saeed Anwar, Salman Khan, Nick Barnes. "Deep Learning-Based Image Captioning with Improved GRU Models." IEEE Access, 2019.
2. Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, Trevor Darrell. "Dense Captioning with Joint Inference and Visual Context." IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.
3. Baoguang Shi, Xiang Bai, Cong Yao. "Scene Text Recognition Using Deep Convolutional Networks." IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.
4. Xinxin Zhu, et al. "Image Captioning and Visual Question Answering Based on Attributes and External Knowledge." IEEE Transactions on Neural Networks and Learning Systems, 2018.
5. Peng Gao, et al. "Hybrid Attention Networks for Visual Question Answering." IEEE Transactions on Image Processing, 2019.
6. Luowei Zhou, et al. "Bidirectional Attentive Fusion with Context Gating for Dense Video Captioning." IEEE Transactions on Image Processing, 2018.
7. Jiasen Lu, et al. "Adaptive Attention Time for Image Captioning." IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.
8. Feng Liu, et al. "Image Captioning with Visual-Semantic LSTM." IEEE Transactions on Multimedia, 2018.
9. Luowei Zhou, et al. "End-to-End Dense Video Captioning with Masked Transformer." IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.
10. Jeffrey Donahue, et al. "Recurrent Convolutional Models for Video Captioning." IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015.
11. Zhi Tian, et al. "Text Proposal Networks for Scene Text Detection." IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016.
12. Zhou Yu, et al. "Co-attention Networks for Visual Question Answering." IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.
13. Zichao Yang, et al. "Stacked Attention Networks for Image Question Answering." IEEE Transactions on Computer Vision and Pattern Recognition, 2016.
14. Karan Sikka, et al. "Text-Attentional Convolutional Neural Networks for Scene Text Detection." IEEE Transactions on Image Processing, 2016.
15. Longyin Wen, et al. "TextSnake: A Flexible Representation for Detecting Text of Arbitrary Shapes." IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.